1/9/1
DIALOG(R)File 351:Derwent WPI

012887342 **Image available**
WPI Acc No: 2000-059176/200005
XRPX Acc No: N00-046401
Search method used in structurizing document e.g. document described in standard general mark-up language format - involves extracting text length of character row included in logic structure, by which adaptation calculation is designated, as search object used for adaptation calculation
Patent Assignee: HITACHI LTD (HITA); KAWASHIMO Y (KAWA-I); MATSUBAYASHI T (MATS-I); OKAMOTO T (OKAM-I); SUGAYA N (SUGA-I); TADA K (TADA-I)
Inventor: KAWASHIMO Y; MATSUBAYASHI T; OKAMOTO T; SUGAYA N; TADA K
Number of Countries: 002 Number of Patents: 003
Patent Family:

Patent No	Kind	Date	Applicat No	Kind	Date	Week	
JP 11316764	A	19991116	JP 98136127	A	19980430	200005	B
US 20020188604	A1	20021212	US 99300594	A	19990428	200301	
			US 2002218495	A	20020815		
US 6496820	B1	20021217	US 99300594	A	19990428	200307	

Priority Applications (No Type Date): JP 98136127 A 19980430
Patent Details:

Patent No	Kind	Lan	Pg	Main IPC	Filing Notes
JP 11316764	A		21	G06F-017/30	
US 20020188604	A1			G06F-007/00	Cont of application US 99300594
US 6496820	B1			G06F-017/30	

Abstract (Basic): JP 11316764 A
NOVELTY - The method involves extracting the text length of a character row included in the logic structure, by which adaptation calculation is designated, as the search object used for adaptation calculation. DETAILED DESCRIPTION - The method begins by searching a document containing the character row designated in a pre-designated logic structure. The adaptation calculation process is then performed in which the adaptation opposing the search conditions designated about the searched document is computed. INDEPENDENT CLAIMS are also included for the following:the search apparatus;and a computer-readable recording medium storing the structurizing document search program.
USE - Used in structurizing document e.g. document described in SGML format.
ADVANTAGE - Structure length of logic structure coinciding with the search conditions during structure designation search can be acquired at high speed. Reduces inaccuracy in searching applicable logic

structure in normal search term. DESCRIPTION OF DRAWING(S) - The figure
shows the component of the search method.
Dwg.1/22
Title Terms: SEARCH; METHOD; STRUCTURE; DOCUMENT; DOCUMENT; DESCRIBE; STANDARD; GENERAL; MARK; UP; LANGUAGE; FORMAT; EXTRACT; TEXT; LENGTH; CHARACTER; ROW; LOGIC; STRUCTURE; ADAPT; CALCULATE; DESIGNATED; SEARCH; OBJECT; ADAPT; CALCULATE
Derwent Class: T01
International Patent Class (Main): G06F-007/00; G06F-017/30
File Segment: EPI
Manual Codes (EPI/S-X): T01-J05B
?

(19)日本国特許庁（ＪＰ）　(12) **公開特許公報（Ａ）**　(11)特許出願公開番号

特開平11－316764

(43)公開日　平成11年(1999)11月16日

(51)Int.Cl.[6]	識別記号	ＦＩ		
Ｇ０６Ｆ　17/30		Ｇ０６Ｆ	15/40	３７０Ａ
				３４０
			15/403	３４０Ａ

審査請求　未請求　請求項の数７　ＦＤ　（全 21 頁）

(21)出願番号　特願平10－136127

(22)出願日　平成10年(1998)４月30日

(71)出願人　000005108
株式会社日立製作所
東京都千代田区神田駿河台四丁目６番地

(72)発明者　多田　勝己
神奈川県川崎市幸区鹿島田890番地　株式会社日立製作所情報・通信開発本部内

(72)発明者　菅谷　奈津子
神奈川県川崎市幸区鹿島田890番地　株式会社日立製作所情報・通信開発本部内

(72)発明者　松林　忠孝
神奈川県川崎市幸区鹿島田890番地　株式会社日立製作所情報・通信開発本部内

(74)代理人　弁理士　笹岡　茂　（外１名）

最終頁に続く

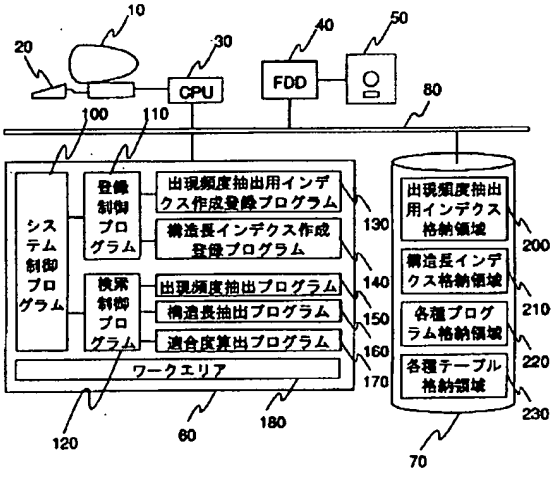
(54)【発明の名称】　構造化文書の検索方法および装置および構造化文書検索プログラムを記録したコンピュータ読み取り可能な記録媒体

(57)【要約】

【課題】　構造化文書を対象として目的とする論理構造を指定する構造指定検索において、検索対象に指定した論理構造のテキスト長を用いた適合度算出処理を高速に実現することにある。

【解決手段】　文書をデータベースに登録する際、指定された論理構造中の検索タームの出現頻度を抽出するための検索用インデクスである出現頻度抽出用インデクスを作成すると共に、登録対象文書中の各文字に対して該当文字に対応する論理構造の識別子と構造長を格納した構造長インデクスを作成し、検索時にはこれらのインデクス群を参照し、その結果得られた出現頻度と構造長を用いて検索結果文書に対する適合度を算出する。

図１　第一の実施例の構成図



【特許請求の範囲】

【請求項１】　予め登録された文書の集合を対象として、指定された論理構造中に指定された文字列を含む文書を検索するステップと、検索結果文書について指定された検索条件に対する適合度を算出する適合度算出ステップを有する文書検索方法において、

前記適合度算出ステップが、適合度算出に用いる検索対象に指定された論理構造に含まれる文字列のテキスト長を抽出する構造長抽出ステップを有することを特徴とする構造化文書の検索方法。

【請求項２】　予め登録された文書の集合を対象として、指定された文字列を含む文書の検索を行なう文書検索方法において、

文書の登録を行なう処理が、登録対象文書に対し、検索時に指定された論理構造中に指定された検索タームを含む文書について該文書を一意に識別するための識別情報と、該論理構造の識別情報と、該論理構造中に検索タームの出現した回数とを抽出するための出現頻度抽出用インデクスを作成登録する出現頻度抽出用インデクス作成登録ステップと、

登録対象文書から抽出した少なくとも１文字以上の部分文字列に対し、該登録対象文書を一意に識別するため識別情報と、該部分文字列に対応する論理構造の識別情報と該論理構造のテキスト長とを格納した構造長インデクスを作成登録する構造長インデクス作成登録ステップを有することを特徴とする構造化文書の検索方法。

【請求項３】　請求項２記載の構造化文書の検索方法において、

文書の検索を行なう処理が、指定された検索タームについて前記登録された出現頻度抽出用インデクスを参照し、該検索タームを指定された論理構造中に含む文書の識別情報と、該論理構造の識別情報と、該論理構造中に検索タームの出現した回数とを抽出する出現頻度抽出ステップと、

指定された検索タームから所定の少なくとも１文字以上の部分文字列を１個以上抽出し、該部分文字列に対し前記登録された構造長インデクスを参照することにより該文字列を含む文書の識別情報と、該文字列の含まれる論理構造の識別情報と、該論理構造の構造長とを抽出する構造長抽出ステップと、

前記出現頻度抽出ステップにより抽出された文書の識別情報、該論理構造の識別情報および該論理構造中に検索タームの出現した回数と、前記論理構造長抽出ステップにより抽出された文書の識別情報、該文字列の含まれる論理構造の識別情報および該論理構造の構造長を用いて検索条件に対する適合度を算出する適合度算出ステップを有することを特徴とする構造化文書の検索方法。

【請求項４】　請求項２記載の構造化文書の検索方法において、

前記出現頻度抽出用インデクス作成登録ステップにおいて作成登録される出現頻度抽出用インデクスは、登録対象文書から所定の部分文字列を抽出し、該部分文字列に対し該文書を一意に識別するための識別情報と、該部分文字列の含まれる該論理構造の識別情報と、該部分文字列の登録対象文書中での位置情報を格納した部分文字列抽出型の出現頻度抽出用インデクスであることを特徴とする構造化文書の検索方法。

【請求項５】　請求項４記載の構造化文書の検索方法において、

検索タームから所定の部分文字列を抽出し、該部分文字列に対し前記登録された部分文字列抽出型の出現頻度情報抽出用インデクスを参照することにより取得した該部分文字列の存在した文書の識別情報、論理構造の識別情報、該文書中での文字位置議をもとに、該検索タームを含む文書の識別情報と、該検索タームの含まれる該論理構造の識別情報と、該論理構造における該検索タームの出現頻度とを抽出する出現頻度抽出ステップと、

指定された検索タームから所定の少なくとも１文字以上の部分文字列を１個以上抽出し、該部分文字列に対し前記登録された構造長インデクスを参照することにより該文字列を含む文書の識別情報と、該文字列の含まれる論理構造の識別情報と、該論理構造の構造長とを抽出する構造長抽出ステップと、

前記出現頻度抽出ステップにより抽出された文書の識別情報、該論理構造の識別情報および該論理構造中に検索タームの出現した回数と、前記論理構造長抽出ステップにより抽出された文書の識別情報、該文字列の含まれる論理構造の識別情報および該論理構造の構造長を用いて検索条件に対する適合度を算出する適合度算出ステップを有することを特徴とする構造化文書の検索方法。

【請求項６】　予め登録された文書の集合を対象として、指定された文字列を含む文書の検索を行なう文書検索装置において、

登録対象文書に対し、検索時に指定された論理構造中に指定された検索タームを含む文書について該文書を一意に識別するための識別情報と、該論理構造の識別情報と、該論理構造中に検索タームの出現した回数とを抽出するための出現頻度抽出用インデクスを作成登録する出現頻度抽出用インデクス作成登録手段と、

登録対象文書から抽出した少なくとも１文字以上の部分文字列に対し、該登録対象文書を一意に識別するため識別情報と、該部分文字列に対応する論理構造の識別情報と該論理構造のテキスト長とを格納した構造長インデクスを作成登録する構造長インデクス作成登録手段と、

指定された検索タームについて前記登録された出現頻度抽出用インデクスを参照し、該検索タームを指定された論理構造中に含む文書の識別情報と、該論理構造の識別情報と、該論理構造中に検索タームの出現した回数とを抽出する出現頻度抽出手段と、

指定された検索タームから所定の少なくとも１文字以上

の部分文字列を１個以上抽出し、該部分文字列に対し前記登録された構造長インデクスを参照することにより該文字列を含む文書の識別情報と、該文字列の含まれる論理構造の識別情報と、該論理構造の構造長とを抽出する構造長抽出手段と、

前記出現頻度抽出手段により抽出された文書の識別情報、該論理構造の識別情報および該論理構造中に検索タームの出現した回数と、前記論理構造長抽出手段により抽出された文書の識別情報、該文字列の含まれる論理構造の識別情報および該論理構造の構造長を用いて検索条件に対する適合度を算出する適合度算出手段を有することを特徴とする構造化文書の検索装置。

【請求項７】　登録対象文書に対し、検索時に指定された論理構造中に指定された検索タームを含む文書について該文書を一意に識別するための識別情報と、該論理構造の識別情報と、該論理構造中に検索タームの出現した回数とを抽出するための出現頻度抽出用インデクスを作成登録する手順と、

登録対象文書から抽出した少なくとも１文字以上の部分文字列に対し、該登録対象文書を一意に識別するため識別情報と、該部分文字列に対応する論理構造の識別情報と該論理構造のテキスト長とを格納した構造長インデクスを作成登録する手順と、

指定された検索タームについて前記登録された出現頻度抽出用インデクスを参照し、該検索タームを指定された論理構造中に含む文書の識別情報と、該論理構造の識別情報と、該論理構造中に検索タームの出現した回数とを抽出する手順と、

指定された検索タームから所定の少なくとも１文字以上の部分文字列を１個以上抽出し、該部分文字列に対し前記登録された構造長インデクスを参照することにより該文字列を含む文書の識別情報と、該文字列の含まれる論理構造の識別情報と、該論理構造の構造長とを抽出する手順と、

前記抽出された文書の識別情報、該論理構造の識別情報および該論理構造中に検索タームの出現した回数と、前記抽出された文書の識別情報、該文字列の含まれる論理構造の識別情報および該論理構造の構造長を用いて検索条件に対する適合度を算出する手順を実行させる構造化文書検索プログラムを記録したコンピュータ読み取り可能な記録媒体。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】SGML(Standard Generalized Markup Language)形式で記述された文書などのように、１件の文書が複数の論理構造で構成される構造化文書に対し、目的とする論理構造だけを対象とした検索を行なう構造指定検索において、検索結果文書に対して検索条件に対する適合度に応じた得点付けを行なう構造化文書の検索方法および装置に関する。

【０００２】

【従来の技術】近年、情報化社会の急速な親展に伴い、ワードプロセッサやパーソナルコンピュータなどを用いて作成される電子化文書情報も爆発的な勢いで増加しつつある。このような状況下で、蓄積された膨大な電子化文書群の中から、必要とする情報を含んだ文書を高速かつ高精度に検索したいという要求が高まっている。このような要求に応える技術として全文検索がある。全文検索では、登録時に登録対象文書中のテキスト全体を計算機システムに入力してデータベース化し、検索時には該当データベース中からユーザの指定した文字列（以下、検索タームと呼ぶ）を含む全ての文書を探し出すことにより、登録時にキーワード付けを行なうことなく、目的とする文書を漏れなく検索することが可能である。

【０００３】しかし、全文検索技術を大規模な文書データベースに対して適用した場合には、以下に示す二つの問題が発生する。まず第一に、検索結果文書中から目的とする文書を探し出すのに時間がかかるという問題が生じる。つまり、大規模な文書データベースを対象として検索を実行した場合には、検索結果として得られる文書の数も膨大なものになる。これらの文書中に、目的とする文書が含まれているか否かを判断するためには、これらの文書全ての内容を読んで理解する必要があり、この処理に膨大な時間を要することになる。また、新たに検索条件を加えることにより検索結果文書を絞り込む方法も考えられるが、この方法では新たに加えた検索条件によって、もとの検索結果中に含まれていた目的とする文書が排除され、検索漏れとなってしまう可能性があるという問題がある。

【０００４】また、第二に全文検索による検索結果には検索ノイズが多く含まれるという問題が生じる。つまり、“検索システム”に関する特許明細書を探す目的で、“検索”という文字列を検索タームに指定して全文検索を実行した場合には、“論理アドレスと物理アドレス間の変換テーブルを検索する”などの言い回しを実施例中に含む“プロセッサ”に関する特許明細書がノイズとして検索されてしまう。

【０００５】これらの問題のうち、検索結果文書から目的とする文書の抽出処理の効率化に対しては、検索結果文書に対し指定された検索条件に対する適合度に応じた得点付けを行い、この得点順に検索結果文書の一覧を表示するスコアリング機能が提案されている。この方法によると、ユーザは得点の高い文書から順に、該当文書が目的の文書であるか否かの判定を行うことができる。また、ある得点以下の文書を判定の対象から外すことにより効率的に検索結果文書の判定を行うことができる。このように、検索結果文書に対し検索条件に対する適合度算出方法の一例が、「Ｉｎｆｏｒｍａｔｉｏｎ　Ｒｅｔｒｉｅｖａｌ」(PRENTICE HALL発行、William B.Frakes, Ricardo Baeza-Yates著)（以下、文献１と呼ぶ）に

示されている。

【０００６】また、第二の問題である検索ノイズの削減に対しては、検索の対象とする論理構造を指定する構造指定検索が提案されている。この方式を用いると、先述した“検索システム”に関する特許明細書を探す場合に「産業上の利用分野」の構造を検索対象に指定し、その中に“検索”という文字列が含まれる明細書だけを抽出することができる。その結果、先述した「実施例」中に“検索”という文字列が含まれるプロセッサに関する特許などはノイズとして検索結果から省くことができる。このように、SGML(ISO 8879:Standard Generalized Markup Language)で記述された文書などのように、１件の文書が複数の論理構造で構成される文書（以下、構造化文書と呼ぶ）に対して、目的とする論理構造だけを対象に指定する構造指定検索を実現する方式の一例として、特願平９－４１８５５号（以下、文献２と呼ぶ）を提案している。

【０００７】以下、文献１と文献２の概略を説明する。文献１では、検索結果の各文書中に指定された検索タームが出現した回数（以下、検索タームの出現頻度と呼ぶ）と各文書のテキスト長を用いて、以下に示す算出式を用いて検索結果文書の適合度算出を行なう方法が記載されている。

ｎｆｒｅｑｉｊ＝（ｌｏｇ2(ｆｒｅｑｉｊ＋１)）／ｌｏｇ2(ｌｅｎｇｔｈｉ)

ただし、ｆｒｅｑｉｊ：検索タームiの文書jにおける出現頻度

ｌｅｎｇｔｈｉ：文書Iのテキスト長

すなわち、検索頻度の出現頻度だけを用いて適合度の算出を行なった場合には、各文書のテキスト長による影響が考慮されないため、検索条件に対する正確な適合度がえられない。つまり、100Bのテキスト中に９個の検索タームを含む文書は、1MBのテキスト中に10個の検索タームを含む文書に比べて、検索タームの出現密度（該当文書中での検索タームの出現確率）の点で高い得点が付けられて然るべきであるにも係わらず、低い得点しか与えられないことになる。この問題を解決するために、文献１では上式に示した通りテキスト長ｌｅｎｇｔｈｉを用いた値で検索タームの出現頻度ｆｒｅｑｉｊの正規化を行なうことにより、精度の高い適合度の算出処理を実現している。

【０００８】次に、文献２に示されている構造指定検索の実現方法について説明する。本方式は、目的とする論理構造だけを検索対象とすることにより、それ以外の論理構造に検索タームが現われる文書を検索結果から除き、全文検索における検索ノイズを低減することを目的としたものである。

【０００９】本方式では、構造化文書をデータベースに登録する際に、登録対象文書の持つ論理構造の解析を行う。そして、文書の登録順に従って各文書の持つ論理構造を順次重ね合わせ、文書中における出現位置および種別が同じである論理構造の要素群および文字列データ群を、それぞれ単一のメタ要素およびメタ文字列として代表させることにより、メタ要素群およびメタ文字列データ群（以下、これらを総称してメタノードと呼ぶ）による木構造データを作成する。そして、これらのメタノードを識別するための一意の識別子（以下、文脈識別子と呼ぶ）を付与することにより、文書データベース中の全文書の論理構造を表わすインデクス（以下、構造インデックスと呼ぶ）を作成する。

【００１０】次に、登録対象文書について該当文書中に含まれる全ての文字列と、前記構造インデックスにおけるメタ文字列データの識別子との対応関係を記録したデータ(以下、構造化全文データ)を生成する。さらに、登録対象文書に関する構造化全文データにおいて、各文字列から所定の部分文字列を抽出し、それらを文書データベース中で識別するための文書識別子、メタ文字列データの文脈識別子および登録対象文書中での文字位置と対応付けたデータ（構造化文字位置情報）として登録することにより検索用のインデクスを生成する。以上が、本文献における一連の登録処理である。

【００１１】そして、検索時には、始めに前記構造インデックスを参照し、検索対象に指定された構造に対応するメタ文字列データの文脈識別子を抽出する。次に、検索タームから所定の部分文字列を抽出し、各部分文字列について検索用のインデクスを参照することにより、検索タームを構成する部分文字列に関する構造化文字位置情報を抽出する。最後に、各部分文字列の構造化文字位置情報について、これらの隣接判定処理を行なう。すなわち、検索タームを構成する各部分文字列の構造化文字位置情報から検索対象に指定した論理構造に対応する文脈識別子を持つものを抽出し、その中で指定された検索タームと同じ部分文字列の並びを持つ文書の文書識別子を抽出することにより構造指定検索を実現している。以上が、文献２における登録処理および検索処理の概要である。

【００１２】次に、本文献における登録処理例について、図を用いて概略の説明をする。本例では、図２に示す構造化文書が登録された場合に、まず論理構造の解析処理を行う。

【００１３】そして、その論理構造を既登録文書における論理構造と重ね合わせることにより、図３に示す構造インデックスを生成する。次に、登録対象文書中の文字列について、図３に示す構造インデックスにおけるメタ文字列データの文脈識別子を対応付けることにより、図４に示す構造化全文データを生成する。さらに、検索用インデクスの生成処理として図４に示す構造化全文データ中の内容文字列から、本文献では隣り合う２文字の文字列を部分文字列として抽出する。そして、各部分文字列に対して該当する文書識別子、文脈識別子および文書

中での文字位置の組を構造化文字位置情報として追記、登録することにより検索用のインデクスを生成する。この結果、例えば“ガー”および“ード”について図5に示すインデクスが生成される。

【００１４】次に、検索時の処理例として“段落”の論理構造中に検索ターム“ガード”が含む文書を検索する際の処理について説明する。検索時には、はじめに図3に示す構造インデックスから、検索対象の論理構造である“段落”に該当する文字列データの文脈識別子として文脈識別子C7,C8,C9,C16,C17,C131を抽出する。次に、検索用インデクスの作成時と同様に、検索ターム“ガード”から隣り合う2文字の文字列として“ガー”および“ード”を抽出する。

【００１５】そして、検索用インデクスから“ガー”および“ード”に関する構造化文字位置情報を抽出し、その中で検索対象構造に該当するメタ文字列データの文脈識別子（本例では，C7,C8,C9,C16,C17,C131）のいずれかに該当するものを取得する。最後に、こうして得られた構造化文字位置情報をもとに、図6に示すように文書識別子および文脈識別子が同一であり、かつ文字位置が隣り合うものを判定することにより、“段落”の論理構造中に検索ターム“ガード”が含まれる文書を検索することが可能になる。

【００１６】

【発明が解決しようとする課題】しかし、文献1における検索結果に対する適合度の算出方式を構造指定検索に適用しようとすると、以下に示す問題が生じる。まず、検索対象に指定した論理構造中の検索タームの出現頻度を正規化するためのテキスト長として文書全体のテキスト長を用いた場合には、他の論理構造に関する文字列およびタグなどの論理構造を記述するための制御用の文字列の影響を受けることになり、正しい適合度を算出することができない。そして、検索対象に指定した論理構造のテキスト長（以下、構造長と呼ぶ）を用いて適合度の算出処理を行なうためには、検索時間が長大化してしまうという問題が生じる。

【００１７】すなわち、図3に示す論理構造を持つ文書データベースを対象として、構造長の取得する手段として、図7に示すように、全登録文書について予め各論理構造のテキスト長を格納した構造長テーブル群を作成する方式が考えられる。しかし、この方法では検索時に検索タームがヒットした文書数分、構造長テーブル群を参照する必要が生じる。この構造長テーブル群は、1エントリを4B、文書の登録件数を100万件とし、文書データベースにおけるメタ要素の数を400とした場合に、1.6GB（＝4B×1,000,000×400）の容量となる。つまり，構造長テーブル群は磁気ディスクなどの2次記憶上に格納されることになり、これをヒットした文書数分アクセスすることになるため検索時間が長大化してしまう。例えば、磁気ディスク上のデータを1回アクセスするのに要する時間を20msとし、検索タームのヒットした文書数を1,000件とすると、構造長テーブル群の参照に20秒（＝20ms×1,000）の時間を要することになる。

【００１８】また、各構造の構造長を図5に示す検索用インデクス中に格納する方式も考えられるが、この方式では、検索用インデクスの容量が図8に示すように膨大化してしまう。つまり、図8において“ガー”の先頭の構造化文字位置情報（文字列データの文脈識別子：C16）は図3からも分かるように“段落”の論理構造（文脈識別子E22）に属するだけでなく“節”の論理構造（文脈識別子E21）にも属している。また、“章”の論理構造（文脈識別子E19）にも属している。このように、該当する構造化文字位置データは“段落”の論理構造（文脈識別子E22）を対象とした検索の場合のみならず、“節”の論理構造（文脈識別子E21）など、上位の論理構造を対象とした検索の場合にも参照される。このため、これらの上位の論理構造の構造長も格納しておく必要がある。また、これらの構造長は検索タームを構成する全ての文字列からも読み出されることになる。つまり、本方式では検索用インデクスの容量が大きくなるだけでなく、検索時に読み出す構造化文字位置文字データの容量の増加につながり、検索に要する時間が著しく長大化するという問題がある。

【００１９】本発明の目的は、適合度算出に用いる論理構造の構造長を高速に取得でし、この論理構造の構造長により適合度算出処理を高速に実現し、精度の高い検索を検索性能を低下させることなく実現することにある。

【００２０】

【課題を解決するための手段】上記目的を達成するため、本発明は、予め登録された文書の集合を対象として、指定された論理構造中に指定された文字列を含む文書を検索するステップと、検索結果文書について指定された検索条件に対する適合度を算出する適合度算出ステップを有する文書検索方法において、前記適合度算出ステップが、適合度算出に用いる検索対象に指定された論理構造に含まれる文字列のテキスト長を抽出する構造長抽出ステップを有するようにしている。

【００２１】また、予め登録された文書の集合を対象として、指定された文字列を含む文書の検索を行なう文書検索方法において、文書の登録を行なう処理が、登録対象文書に対し、検索時に指定された論理構造中に指定された検索タームを含む文書について該文書を一意に識別するための識別情報と、該論理構造の識別情報と、該論理構造中に検索タームの出現した回数とを抽出するための出現頻度抽出用インデクスを作成登録する出現頻度抽出用インデクス作成登録ステップと、登録対象文書から抽出した少なくとも1文字以上の部分文字列に対し、該登録対象文書を一意に識別するため識別情報と、該部分文字列に対応する論理構造の識別情報と該論理構造のテキスト長とを格納した構造長インデクスを作成登録する

構造長インデクス作成登録ステップを有するようにしている。

【００２２】また、文書の検索を行なう処理が、指定された検索タームについて前記登録された出現頻度抽出用インデクスを参照し、該検索タームを指定された論理構造中に含む文書の識別情報と、該論理構造の識別情報と、該論理構造中に検索タームの出現した回数とを抽出する出現頻度抽出ステップと、指定された検索タームから所定の少なくとも１文字以上の部分文字列を１個以上抽出し、該部分文字列に対し前記登録された構造長インデクスを参照することにより該文字列を含む文書の識別情報と、該文字列の含まれる論理構造の識別情報と、該論理構造の構造長とを抽出する構造長抽出ステップと、前記出現頻度抽出ステップにより抽出された文書の識別情報、該論理構造の識別情報および該論理構造中に検索タームの出現した回数と、前記論理構造長抽出ステップにより抽出された文書の識別情報、該文字列の含まれる論理構造の識別情報および該論理構造の構造長を用いて検索条件に対する適合度を算出する適合度算出ステップを有するようにしている。

【００２３】また、前記出現頻度抽出用インデクス作成登録ステップにおいて作成登録される出現頻度抽出用インデクスは、登録対象文書から所定の部分文字列を抽出し、該部分文字列に対し該文書を一意に識別するための識別情報と、該部分文字列の含まれる該論理構造の識別情報と、該部分文字列の登録対象文書中での位置情報を格納した部分文字列抽出型の出現頻度抽出用インデクスであるようにしている。

【００２４】また、検索タームから所定の部分文字列を抽出し、該部分文字列に対し前記登録された部分文字列抽出型の出現頻度情報抽出用インデクスを参照することにより取得した該部分文字列の存在した文書の識別情報、論理構造の識別情報、該文書中での文字位置議をもとに、該検索タームを含む文書の識別情報と、該検索タームの含まれる該論理構造の識別情報と、該論理構造における該検索タームの出現頻度とを抽出する出現頻度抽出ステップと、指定された検索タームから所定の少なくとも１文字以上の部分文字列を１個以上抽出し、該部分文字列に対し前記登録された構造長インデクスを参照することにより該文字列を含む文書の識別情報と、該文字列の含まれる論理構造の識別情報と、該論理構造の構造長とを抽出する構造長抽出ステップと、前記出現頻度抽出ステップにより抽出された文書の識別情報、該論理構造の識別情報および該論理構造中に検索タームの出現した回数と、前記論理構造長抽出ステップにより抽出された文書の識別情報、該文字列の含まれる論理構造の識別情報および該論理構造の構造長を用いて検索条件に対する適合度を算出する適合度算出ステップを有するようにしている。

【００２５】また、予め登録された文書の集合を対象として、指定された文字列を含む文書の検索を行なう文書検索装置において、登録対象文書に対し、検索時に指定された論理構造中に指定された検索タームを含む文書について該文書を一意に識別するための識別情報と、該論理構造の識別情報と、該論理構造中に検索タームの出現した回数とを抽出するための出現頻度抽出用インデクスを作成登録する出現頻度抽出用インデクス作成登録手段と、登録対象文書から抽出した少なくとも１文字以上の部分文字列に対し、該登録対象文書を一意に識別するため識別情報と、該部分文字列に対応する論理構造の識別情報と該論理構造のテキスト長とを格納した構造長インデクスを作成登録する構造長インデクス作成登録手段と、指定された検索タームについて前記登録された出現頻度抽出用インデクスを参照し、該検索タームを指定された論理構造中に含む文書の識別情報と、該論理構造の識別情報と、該論理構造中に検索タームの出現した回数とを抽出する出現頻度抽出手段と、指定された検索タームから所定の少なくとも１文字以上の部分文字列を１個以上抽出し、該部分文字列に対し前記登録された構造長インデクスを参照することにより該文字列を含む文書の識別情報と、該文字列の含まれる論理構造の識別情報と、該論理構造の構造長とを抽出する構造長抽出手段と、前記出現頻度抽出手段により抽出された文書の識別情報、該論理構造の識別情報および該論理構造中に検索タームの出現した回数と、前記論理構造長抽出手段により抽出された文書の識別情報、該文字列の含まれる論理構造の識別情報および該論理構造の構造長を用いて検索条件に対する適合度を算出する適合度算出手段を有するようにしている。

【００２６】また、構造化文書検索プログラムを記録したコンピュータ読み取り可能な記録媒体であり、登録対象文書に対し、検索時に指定された論理構造中に指定された検索タームを含む文書について該文書を一意に識別するための識別情報と、該論理構造の識別情報と、該論理構造中に検索タームの出現した回数とを抽出するための出現頻度抽出用インデクスを作成登録する手順と、登録対象文書から抽出した少なくとも１文字以上の部分文字列に対し、該登録対象文書を一意に識別するため識別情報と、該部分文字列に対応する論理構造の識別情報と該論理構造のテキスト長とを格納した構造長インデクスを作成登録する手順と、指定された検索タームについて前記登録された出現頻度抽出用インデクスを参照し、該検索タームを指定された論理構造中に含む文書の識別情報と、該論理構造の識別情報と、該論理構造中に検索タームの出現した回数とを抽出する手順と、指定された検索タームから所定の少なくとも１文字以上の部分文字列を１個以上抽出し、該部分文字列に対し前記登録された構造長インデクスを参照することにより該文字列を含む文書の識別情報と、該文字列の含まれる論理構造の識別情報と、該論理構造の構造長とを抽出する手順と、前記

抽出された文書の識別情報、該論理構造の識別情報および該論理構造中に検索タームの出現した回数と、前記抽出された文書の識別情報、該文字列の含まれる論理構造の識別情報および該論理構造の構造長を用いて検索条件に対する適合度を算出する手順を実行させるようにしている。

【００２７】

【発明の実施の形態】本発明の適合度算出機能を備えた構造化文書検索システムの第一の実施例を図１に示す。本図に示す構造化文書検索システムは検索結果を表示するディスプレイ１０、登録および検索のコマンドを入力するキーボード２０、登録処理および検索処理を実行する中央演算処理装置ＣＰＵ３０、フロッピディスクからデータを読み出すフロッピディスクドライバ４０、データベースへ登録する構造化文書データを格納したフロッピディスク５０、登録および検索用のプログラムならびにデータなどを一時的に格納する主メモリ６０、各種データおよびプログラムを格納する磁気ディスク７０およびこれらを接続するバス８０で構成される。主メモリ６０にはシステム制御プログラム１００、登録制御プログラム１１０、検索制御プログラム１２０、出現頻度抽出用インデクス作成登録プログラム１３０、構造長インデクス作成登録プログラム１４０、出現頻度抽出プログラム１５０、構造長抽出プログラム１６０、適合度算出プログラム１７０が磁気ディスク７０から読み出されるとともに、ワークエリア１８０が確保される。また、磁気ディスク７０には出現頻度抽出用インデクス格納領域２００、構造長インデクス格納領域２１０、各種プログラム格納領域２２０および各種テーブル格納領域２３０が確保されている。なお、本実施例ではこれらの格納領域を磁気ディスク上７０上に確保したが、光磁気ディスク装置など他の二次記憶装置であっても構わない。以上が本構造化文書検索システムの構成である。

【００２８】次に、本実施例に示す構造化文書検索システムの文書登録時の処理の概要について説明する。本実施例では、検索対象とする論理構造を識別する方法として、文書登録の前に予め論理構造の型定義文を解析することにより繰り返しを持つ論理構造を抽出し、その繰り返し回数に上限値を設定することにより、各論理構造を一意に識別するための識別子（構造識別子）および各構造に対する文字列の識別子を固定的に割り振る方式について説明する。

【００２９】まず、本実施例では、文書の登録前に事前に登録対象となる文書の論理構造を解析し、繰り返しのある論理構造を抽出しておく。すなわち、例えば、図２に示す論理構造の文書においては、図１０（なお、本図において繰り返しを持つ論理構造に対しては２重の枠線で示している）に示す通り“執筆者”中の“氏名”、“本文”中の“章”、さらに“文献リスト”の中の“文献”が繰り返し構造として定義されている。また、“章”中の“段落”、“節”および“備考”、さらに“節”の下の“段落”、“項”および“備考”、そして“項”の下の“段落”および“備考”が繰り返し構造であり、“文献リスト”の下位構造である執筆者が繰り返し構造が繰り返し構造として定義されている。このような繰り返し構造は構造化文書における論理構造の定義文(例えば、本実施例に示すＳＧＭＬ文書においては文書型定義ＤＴＤ(Document Type Definition、図９にその例を示す)を参照することにより抽出することができる。

【００３０】そして、これらの繰り返し構造に対し、繰り返し数の最大値を定義として１０を与えた場合の構造の識別子ならびに各構造に対する文字列の識別子の割り当て方式について説明する。まず、図１０に示す構造において、“論文”に対し構造識別子の初期値であるE1を割り当てる。そして、最初に出現する“タイトル”の論理構造に対して構造の識別子としてE2を割り当て、“タイトル”の構造には文字列データが格納されるため、対応する文字列の識別子としてC1を割り当てる。次に“執筆者”の構造に着目し、“執筆者”の論理構造に対し構造の識別子E3を割り当て、その下位構造である、“執筆者”に着目する。この“執筆者”の論理構造には、繰り返しを持つ“名前”の構造が定義されているため、これらに対し１０個の構造識別子（E4～E13まで）を割り当て、各構造に関する文字列の識別子としてC2～C11を割り当てる。さらに、その後の現れる“日付”の論理構造に対して構造の識別子としてE14を、文字列に対する識別子としてC12を割り当てる。

【００３１】そして、次に“本文”の論理構造に着目する。ここでは、まず始めに“本文”の構造に対して構造の識別子E15を割り当て“本文”の下位の論理構造である“章”に着目する。ここで“章”は最大10回の繰り返し回数を持つ構造と定義されるため、“1章”から“10章”にかけて10個の構造識別子E16～E25を割り当てる。また、“章”の下位構造である“章題”についても同様に“1章”の章題から“10章”の章題に対し、それぞれ１個ずつ、合計10個の構造識別子E26～E35と各構造に対する文字列の識別子としてC13～C22を割り当てる。さらに、“段落”については“1章”から“10章”の各章に対し、それぞれ“段落1”から“段落10”までの10個の構造識別子を、すなわち100個の構造識別子E36～E135を割り当てるとともに、各構造に対し文字列の識別子C23～C122を割り当てる。

【００３２】そして、引き続き“節”の論理構造に着目する。“節”についても“1章”から“10章”の各章について最大10節までが定義できることから、“1章1節”から“10章10節”にかけて個別に構造の識別子を割り当てることにより、合計100個の構造識別子E136～E235を割り当てる。以下、備考ならびに各節における“節題”、“段落”、“項”および“節”の構造中の“備

考”に着目し、上記と同様の処理で構造識別子および文字列の識別子を割り当てていくことにより、各論理構造および文字列を一意に識別するための識別子を割り当てておく。

【００３３】また、これらの論理構造および文字列を一意に識別するための識別子を、例えば図１１に示すデータ形式で論理構造の管理テーブルとして格納しておく。なお、本図において矢印はポインタの差す値を表し、テーブルデータ中の“rep”は繰り返し構造を持つことを表す特殊コードを示す。また“rep”の右側の値“10”は、繰り返しを持つ該当構造の最大繰り返し数が10であることを示している。また、各構造が最下位の構造であるか、次の階層へのポインタ情報であるかは、各テーブルの中の各エントリ値が文字列の識別子を表すC1、C2…の系列値であるか、ノードへのポインタを表すptr1、ptr2、…の系列値であるか否かにより識別することができる。以上が、本実施例における文書登録の前処理の内容である。

【００３４】次に、本実施例における文書登録時の処理について説明する。キーボード２０から文書の登録コマンドが入力されると、システム制御プログラム１００は登録制御プログラム１１０を起動し、図１２に示す文書の登録処理を開始する。登録制御プログラム１１０は、フロッピディスク５０に格納されている全ての登録対象文書について、ステップ１００１からステップ１００４までに示す一連の処理を繰り返し実行する（ステップ１０００）。

【００３５】まず、ステップ１００１ではフロッピディスクドライバ４０を通じてフロッピディスク５０に格納されている登録対象文書群から未処理の文書を１個選択し、主メモリ６０上のワークエリア１８０に読み出す。次に、ステップ１００２で、ステップ１００１で読み込んだ登録対象文書に対し、文書データベース中で該当文書を一意に識別するための番号である文書識別子を割り当てる。

【００３６】さらに、ステップ１００３において主メモリ６０上の登録対象文書に対し出現頻度抽出用インデクス作成登録プログラム１３０を実行し、登録対象文書中の全ての文字列に対し、該当文字列が含まれる論理構造の識別子との対応関係を示した情報（構造化全文データ）を主メモリ６０上のワークエリア１８０に格納する。そして、構造化全文データ中の文字列から全ての１文字および互いに隣り合う２文字の文字列を抽出し、それらの文字および文字列に対し検索用のインデクスを生成し、磁気ディスク７０上の出現頻度抽出用インデクス２００を追加し、更新する。

【００３７】最後に、ステップ１００４において、主メモリ６０上に格納された登録対象文書中の文字列と該当文字列が含まれる論理構造の識別子の対応関係を示した構造化全文データを入力として、構造長インデクス作成登録プログラム１４０を実行する。そして、登録対象文書中に出現した文字について、該当する文書識別子と各文字の出現した論理構造の識別番号と該当論理構造の構造長を組にして、磁気ディスク７０上に格納した構造長インデクス２１０に追記、更新する。以上が本実施例における登録処理の概要である。

【００３８】次に、図１２におけるステップ１００３とステップ１００４の詳細、すなわち本実施例における出現頻度抽出用インデクス作成登録プログラム１３０の処理手順および構造長インデクス作成登録プログラム１４０の処理手順について説明する。

【００３９】まず、第一に、ステップ１００３における出現頻度抽出用インデクス作成登録プログラム１３０の処理手順を図１３に示すPADを用いて説明する。出現頻度抽出用インデクス作成登録プログラム１３０では、ステップ１１００で、図１１に示す構造識別子管理テーブルを参照しながら登録対象テキストの解析処理を行う。具体的には、図２に示す登録文書中の構造名（“<;”ないしは“<;/”と“>;”の間で区切られた文字列）と図１１に示す構造識別子管理テーブル中の論理構造名を照らしあわせながら、登録文書中の論理構造を辿ることにより、各文字列に対する構造識別子を抽出し、図１４に示す構造化全文データを生成する。

【００４０】次にステップ１１０１で構造化全文データにおけるテキスト（内容文字列）から全ての１文字および互いに隣り合う２文字の文字列を抽出する。具体的には、例えば、図１４に示す構造化全文データ中のタイトルの構造（文字列の構造識別子：C1）に該当する内容文字列「SGML文書変換言語の開発とその適用事例」から“Ｓ”、“SG”、“G”、“GM”、“M”、“ML”、“L”、“L文”、“文”、“文書”、・・・などを抽出する。以下、同様にほかの論理構造中の内容文字列からも全ての１文字および互いに隣り合う２文字の文字列を抽出する。そして、ステップ１１０２においてステップ１１０１で抽出した文字および文字列を木構造データとして登録するとともに、該当文書の識別子と各文字列の属する論理構造の識別子と各文字および文字列の出現した文字位置（２文字の文字列については、その前方の文字の出現した位置）と併せてインデクスデータとして格納する。

【００４１】すなわち、図１４に示した構造化全文データにおいて“S”という文字列はC1(タイトル)の論理構造の１文字目に出現していることから図１５における“S”に該当するインデクス(IDX1)の１番目のエントリに文書識別子D1とともに、文字列の構造識別子C1と文字位置“１”を格納する。また、“SG”についても同様に文書識別子D1、文字列の構造識別子C1と文字位置“１”を組みにして“SG”に該当するインデクス(IDX8)の１番目のエントリに格納する。以下同様に、登録処理を繰り返していく。

【００４２】さらに、“S”はC23(章1-段落1)の論理構造8文字目およびC24(章1-段落2)の論理構造5文字目に出現していることから、これらのデータを図２２における“S”に該当するインデクス(IDX1)の２番目および３番目のエントリに格納していく。以上が、本実施例における出現頻度抽出用インデクス作成登録プログラム１３０の処理内容である。

【００４３】引き続き、図１２におけるステップ１００４の詳細、すなわち本実施例における構造長インデクス作成登録プログラム１４０の処理手順について図１６に示すPADを用いて説明する。はじめに、構造長作成登録プログラム１００４はステップ１２００で、登録対象文書（図１７の例により後述）における各論理構造に現われた文字の出現情報を記録するための構造別文字成分表および各論理構造の構造長を算出するための構造長リスト（図１７の例により後述）の格納領域を主メモリ６０上のワークエリア１８０にアロケートする。また、初期設定として構造別文字成分表および構造長リストの各エントリに‘０’を設定する。

【００４４】次に、ステップ１２０１で登録対象文書に対応する構造化全文データにおける全ての内容文字列に対しステップ１２０２からステップ１２０６までの一連の処理を実行する。まず、ステップ１２０２では該当する内容文字列の属する、上位構造を含む全ての論理構造について構造の識別子を取得する。そして、ステップ１２０３で該当する内容文字列中の全ての文字列に対して、１文字の抽出（ステップ１２０４）、構造別文字成分表の該当文字に対応するエントリに対してステップ１２０２で取得した構造の識別子に対応するビットに‘１’を設定し（ステップ１２０５）、構造長リストにおけるステップ１２０２で取得した構造の識別子に対応する値に１を加算することにより構造長データを更新する（ステップ１２０６）。以上の処理を内容文字列の末尾まで繰り返すことにより、各論理構造の構造長および各論理構造における各文字の出現情報を記録する。

【００４５】以上の処理により作成した構造別文字成分表に対して、ステップ１２０７において全ての文字コードに対応するエントリについて以下の処理を行う。すなわち、構造別文字成分表の各文字コードのエントリに着目し、‘１’が設定されているビットが存在するか否かを判定し（ステップ１２０８）、‘１’が設定されているビットが存在する場合には該当論理構造に対応する構造の識別子を格納するとともに、該当構造識別子に対応する構造長リストのデータを参照することにより取得し、磁気ディスク７０上の構造長インデクス格納領域２１０の該当文字のデータ末尾に追記する（ステップ１２０９）。以上が、本実施例における構造長インデクス作成登録プログラム１４０の処理内容である。

【００４６】さらに、図１４に示す構造化全文データが登録された時の本プログラムの処理例について例を挙げて説明する。ステップ１２０１では、図１７に示す構成で構造別文字成分表および構造長リストの格納領域のアロケートおよび初期設定を行う。次に、ステップ１２０２における繰り返し処理では、まずはじめに図１４における構造化全文データにおける第一行目の内容文字列（構造識別子C1）に着目する。そして、ステップ１２０３では、図１１に示す構造識別子管理テーブルを上位から探索し構造識別子C1を抽出することにより、構造識別子C1に対応する内容文字列を含む論理構造の識別子としてE1およびE2を取得する。そして、ステップ１２０３では内容文字列“SGML文書変換言語の開発とその適用事例”に着目し、ステップ１２０４では先頭文字である“S”を抽出する。そして、ステップ１２０５で図１７に示した構造別文字成分表の文字コード“S”のエントリにおける構造の識別子E1とE2に該当するビットに‘１’を設定する。そして、ステップ１２０６で構造長リストにおけるE1とE2における値にそれぞれ１を加算することにより、E1とE2に対する値に‘１’が設定されることになる。

【００４７】次に、ステップ１２０３では次の文字として“G”を抽出し、ステップ１２０５で図１７に示した構造別文字成分表の文字コード“G”のエントリにおける構造識別子E1とE2に該当するビットに‘１’を設定する。そして、ステップ１１０６で構造長リストにおけるE1とE2における値にそれぞれ１を加算することにより、E1とE2の値は‘２’となる。以下、同様の処理を“M”、“L”、“文”、“書”、・・・について繰り返す。そして、識別子C1に対応する内容文字列“SGML文書変換言語の開発とその適用事例”について処理が終了すると、次の内容文字列“神奈川一郎”に着目し、ステップ１２０２以下同様の処理を繰り返す。以上の処理を図１４に示す構造化全文データ全体に繰り返すことにより、図１８に示す構造別文字成分表および構造長リストが生成されることになる。

【００４８】次に、ステップ１２０７における繰り返し処理では構造別文字成分表（図１８）における各文字コードに対応するエントリに着目する。すなわち、まずはじめに図１８における構造別文字成分表の“a”に対応するエントリに着目し、ステップ１２０８で‘１’が設定されているビットが存在するか否かを判定する。そして“a”については‘１’が設定されているビットが存在しないため、ステップ１２０９を実行することなく次の文字コードに対応するエントリに着目する。そして、例えば“G”のように‘１’が設定されているビットが存在する場合には、ステップ１２０９で‘１’が設定されている論理構造の識別子としてE1,E2,E8,E9,E11およびE12を抽出する。そして、それぞれの構造の識別子について構造長リストを参照することにより構造長を取得する。こうして得られた構造の識別子と構造長の組（E1と9,988，E2と20，E8と8,224，・・・）を文書識別子(D1)

と合わせて文字コード別に格納することにより図１９に示す構造長インデクスを生成する。以上が、本実施例における登録処理内容である。

【００４９】なお、本実施例では登録対象１件毎に磁気ディスク７０上の出現頻度抽出用インデクス２００および構造長インデクス２１０を更新する方式について述べたが、全ての登録対象文書に対する出現頻度抽出用インデクス情報および構造長インデクス情報を、主メモリ６０上のワークエリア１８０に作成したあと、これらを一括して磁気ディスク７０上の出現頻度抽出用インデクス２００および構造長インデクス２１０を更新する方式であっても構わない。

【００５０】次に、検索時の処理について説明する。本発明におけるドキュメント管理システムに対してネットワークを介してユーザから検索コマンドが入力されると、システム制御プログラム１００は検索制御プログラム１２０を起動し、文書の検索処理を開始する。

【００５１】文書検索時の処理を図２０に示すPADを用いて説明する。始めに、検索制御プログラム１２０はステップ２０００で出現頻度抽出プログラム１５０を起動する。出現頻度抽出プログラム１５０では、ユーザの指定した検索条件で磁気ディスク７０上の出現頻度抽出用インデクス格納領域２００に格納された出現頻度抽出用インデクスあるいはこの内の一部または全部を主メモリ６０上のワークエリア１８０に読み出したコピーを参照し、指定された論理構造中に指定された検索タームが含まれる文書の識別子、検索タームを含む論理構造の識別子および検索タームの出現頻度を取得し、主メモリ６０上のワークエリア１９０内に格納する。

【００５２】次に、検索制御プログラム１２０はステップ２００１で構造長抽出プログラム１６０を起動し、登録時に作成し磁気ディスク７０上の出現頻度抽出用インデクス格納領域２００に格納した構造長インデクスあるいはこの内の一部または全部を主メモリ６０上のワークエリア１８０に読み出したコピーを参照し、検索タームの含まれる論理構造に関する構造長を取得し、ワークエリア１８０内に格納する。

【００５３】最後に、検索制御プログラム１２０はステップ２００２で適合度算出プログラム１７０を起動する。適合度算出プログラム１７０では、出現頻度抽出プログラム１５０により得られた文書識別子、検索タームを含む論理構造の識別子および検索タームの出現頻度と、構造長抽出プログラム１６０により得られた検索タームの含まれる論理構造に関する構造長を用いて、検索条件に対する適合度を算出する。

【００５４】これを検索結果文書の一覧情報の一部として付加してユーザに返送し検索制御プログラム１２０を終了する。なお、本処理のおける検索条件に対する適合度の算出方法は、公知例１に開示してある算出式を用いて算出した結果であっても構わない。以上が検索時の処理の概要である。

【００５５】次に、図２０におけるステップ２０００およびステップ２００１の詳細、すなわち本実施例における出現頻度抽出プログラム１５０および構造長抽出プログラム１６０の処理手順について用いて説明する。

【００５６】まず始めに、出現頻度抽出プログラム１５０では図２１に示すPAD図におけるステップ２１００において、図１１に示す構造識別子の管理テーブルを参照し、指定した論理構造に対応する文字列の構造識別子を抽出する。次に、ステップ２１０１において指定された検索タームの文字列をキーに図１５に示す出現頻度抽出用インデクスの木構造データ部を探索することにより、部分文字列に展開する。そしてステップ２１０２において出現頻度抽出用インデクス２００を参照し、ステップ２１０１で抽出した部分文字列に関するインデクスデータを読み出し、インデクス間の隣接判定処理を行うことにより指定された検索タームが指定された論理構造中に含まれる文書の識別子、構造の識別子、および検索タームの出現頻度を抽出、処理を終了する。以上が、本実施例における出現頻度抽出プログラム１５０の処理の概要である。

【００５７】引き続き、構造長抽出プログラム１６０の処理内容について図２２に示すPADを用いて説明する。まず始めに、ステップ２２００において検索タームの先頭一文字を抽出してくる。そしてステップ２２０１では、磁気ディスク７０上の構造長インデクス格納領域２００に格納された構造長インデクス、または予め主メモリ６０上のワークエリア１８０に読み出された構造長インデクスの一部または全体のコピーから、ステップ２２００において抽出した文字に関する情報を抽出することにより、検索タームの先頭に位置する文字を含む文書識別子、論理構造の識別子および該当構造の構造長を抽出し主メモリ６０上のワークエリア１８０に読み込む。最後にステップ２２０２では、ステップ２２０１で読み込まれた文書識別子、論理構造の識別子および該当構造の構造長のうち、検索対象に指定された論理構造に関する情報のみを主メモリ６０上のワークエリア１９０内の別領域にコピーする。以上が、本実施例における構造長抽出プログラム１６０の処理内容である。

【００５８】これまで示した、本実施例における検索プログラムの処理内容の詳細について、図１１に示す論理構造の文書データベースに対し、タイトルの論理構造に“SGML”という検索タームを含む文書の検索という条件を指定した場合について具体的に例を挙げて説明する。まず、図２１におけるステップ２１００において図１１に示す構造識別子の管理テーブルを参照し、検索対象に指定された論理構造である“論文”の下の“タイトル”の論理構造に着目し、該当論理構造の識別子であるE2を抽出する。そして、該当構造に含まれる全ての文字列に関する構造識別子を抽出してくる。本例では、E2の構造

は最下位の論理構造であり、該当する文字列に関する構造識別子としてはC1が抽出されることになる。そして、ステップ２１０１において、検索タームである“SGML”という文字の並びで図１５に示す出現頻度抽出用インデクスの木構造データ部を探索することにより、検索タームを構成する部分文字列として“SG”と“ML”を抽出する。そして、インデクス格納部から該当するインデクス(IDX8およびIDX10)を抽出する。そして、これらのインデクスから、検索対象構造である構造識別子C1に該当するものだけを抽出し、“SG”と“ML”のインデクスが同一の文書識別子であり、同一の構造識別子C1であり、かつ文字位置が２文字離れて隣接するものを抽出する。本例では検索条件を満たす文書として文書識別子D1、構造識別子としてC1、また出現頻度として“１”を抽出し、この検索結果をワークエリア１８０に格納する。

【００５９】次に、構造長抽出プログラム１６０では検索ターム“SGML”の先頭文字である“S”に着目し、構造長インデクス（図１９）を参照し、“S”を含む論理構造の文書識別子および構造長を取得する。構造長を取得しワークエリア１８０に格納する。そして、適合度算出プログラム１７０では、ワークエリア１８０に格納された“S”に関する構造長のうち検索対象構造の識別子であるC1に対応するものを抽出し、検索タームの出現頻度と合わせて各論理構造における検索結果の適合度算出を行う。最後に、検索制御プログラム１２０は、以上の処理によって各論理構造毎の検索条件に対する適合度を受け取ると、これをシステム制御プログラム１００を介して検索者に返送することにより検索処理を終了する。以上が本実施例における文書検索時の処理内容である。

【００６０】なお、本実施例に示した文書の検索処理における構造長インデクス作成登録ステップ１４０において、図１６におけるステップ１２０６の構造長リスト値の加算処理では内容文字列から抽出した１文字に対し常に‘１’を加算することにより、構造長として文字数を算出する方式について説明した。しかし、この加算処理において、内容文字列から抽出した各文字のバイト長（例えば、１バイト文字については‘１’、２バイト文字については‘２’）を加算することにより、構造長として容量を算出することも可能である。

【００６１】このように、本発明によると構造指定検索時に検索条件に合致した論理構造を高速に取得することが可能になり、検索対象に指定された論理構造における検索タームを該当論理構造の構造長で正規化した精度の高い検索を高速に実現できるようになる。なお、本発明における構造指定検索における適合度算出方式では、構造長の取得時に読み込む構造長インデクスの容量は約80kB(構造識別子および構造長を4Bのデータとして扱い、10万件の文書データベースを対象として検索ターム先頭文字の出現頻度確率を１％、また1文書中に検索タームを含む論理構造が平均で10個存在する条件を仮定)であり、大規模な文書データベースに対しても検索のレスポンスをほとんど劣化させることはない。

【００６２】なお、本実施例では構造長インデクスの作成時に登録対象文書から全ての１文字を抽出し、検索時には検索タームから先頭の１文字を抽出する方式について述べた。しかし、検索タームを構成する任意の文字に関する構造長インデクスを参照することができることは明らかである。また、文書の登録時に各文字に対応する構造長インデクスデータの容量をテーブルとして格納しておき、検索時には、検索タームに含まれる文字に対し該当テーブルを参照し、構造長インデクス容量の小さい文字に関するデータを読み込むことにより、構造長インデクスを読み込む時間を短縮することも可能である。

【００６３】さらに、本実施例では登録対象文書から１文字を抽出し、これを構造長インデクスに登録し、検索時にも検索タームから１文字を抽出し、これに対し構造長インデクスを参照する方式について述べた。しかし、登録対象文書中の２文字以上の文字列に対しても、同様の処理により構造長インデクスへの登録を行い、検索時に検索タームから最長の部分文字列を抽出する方式であっても構わない。この方式により、構造長インデクスの容量は増大し、データ登録に必要となる磁気ディスクの容量の増加が考えられるが、検索時に読み込む構造長インデクスの容量を削減することができ、ひいては構造長の抽出処理をさらに短縮することが可能になる。

【００６４】また、本実施例では登録対象文書中から抽出した１文字に対して、文書識別子と該当文字を含む全ての論理構造の識別子および構造長を格納しておき、検索時には検索対象に指定されなかった論理構造に対する識別子および構造長も含めた形で構造長を抽出し、適合度算出ステップにおいて検索タームの存在した論理構造に関する構造長のみを適合度算出に使用する方式について述べた。しかし、構造長インデクス作成時に各論理構造の構造長を論理構造毎に別々の領域に格納し、検索時には検索タームから抽出した部分文字列について、検索対象に指定された論理構造に関する構造長インデクスを参照する方式を採ることによって、検索時に読み込む構造長インデクスの容量を削減することができ、ひいては構造長の抽出処理をさらに短縮することが可能になる。

【００６５】さらに、本実施例における出現頻度抽出用インデクス作成登録プログラム１３０、出現頻度抽出用プログラム１５０および出現頻度抽出用インデクス２００において、検索対象とする論理構造を識別するための識別子の付与する方法として、繰り返し構造を持つ論理構造において、繰り返しの最大数（１０）を定義することにより、固定的に各論理構造を識別する識別子を付与する方法について説明した。しかし、構造内での繰り返し数の上限値を各構造に対し個別に設定することも可能である。また、公知例２において開示されているように、各登録文書において出現した論理構造を重ねあわせ

ていくことにより、無駄な構造の識別子を割り振ることなく構造の上限値にとらわれないように構造の識別子を管理する方法であっても構わない。

【００６６】最後に、本実施例における出現頻度抽出用インデクス作成登録プログラム１３０、出現頻度抽出用プログラム１５０および出現頻度抽出用インデクス２００において、検索タームの出現した頻度を抽出するインデクスの作成方法としては、登録文書中の１文字および隣り合う２文字の文字列に対してインデクスを作成する方法について述べたが、その他の検索アルゴリズムとして公知例２において示されているように、隣り合う２文字の文字列だけに対しインデクスを作成する方法であっても構わない。また、１文字以上の部分文字列（２文字、３文字…を含む）、辞書や形態素解析ないしは登録文書中での出現頻度などの統計情報により抽出した単語等のうち少なくとも１つ以上に対してインデクスを作成する方法であっても構わない。さらに、オートマトンを用いた文字列照合アルゴリズムを適用した方法などであっても構わない。

【００６７】

【発明の効果】本発明によると、予め作成した構造長インデクスを参照して検索タームに関する論理構造の構造長を取得することにより、構造指定検索時に検索条件に合致した論理構造の構造長を高速に取得することが可能になり、ひいては検索対象に指定された論理構造における検索タームを該当論理構造の構造長で正規化した精度の高い検索を検索性能を低下させることなく実現することが可能になる。

【図面の簡単な説明】

【図１】本発明の実施例における構成を示す図である。

【図２】ＳＧＭＬ形式で記述された構造化文書の例を示す図である。

【図３】図２に示した構造化文書の論理構造を示す図である。

【図４】公知例２における構造化全文データのデータ形式を示す図である。

【図５】公知例２における検索用インデクスの構成を示す図である。

【図６】公知例２における検索処理例を示す図である。

【図７】論理構造毎に、各文書識別子に対応する構造長を構造長テーブルとして格納する方式例の概略を示す図である。

【図８】公知例２において、各構造の構造長を検索用インデクス内に格納する方式の概略を示す図である。

【図９】構造化文書（SGML）における文書の型定義（DTD）の例を示す図である。

【図１０】図２に示す構造化文書の論理構造を示す図である。

【図１１】本発明の実施例における構造識別子管理テーブルの構成を示す図である。

【図１２】本発明の実施例における文書登録処理フローを示す図である。

【図１３】本発明の実施例における出現頻度抽出用インデクス作成登録プログラムの処理内容を示す図である。

【図１４】本発明の実施例における構造化全文データの例を示す図である。

【図１５】本発明の実施例における出現頻度抽出用インデクスの内容を示す図である。

【図１６】本発明の実施例における構造長インデクス作成登録プログラムの処理フローを示す図である。

【図１７】本発明の実施例における構造別文字成分表および構造長リストの構成を示す図である。

【図１８】本発明の実施例における構造別文字成分表および構造長リストの例を示す図である。

【図１９】本発明の実施例における構造長インデクスの構成を示す図である。

【図２０】本発明の実施例における検索時の処理内容を示す図である。

【図２１】本発明の実施例における出現頻度抽出プログラムの処理フローを示す図である。

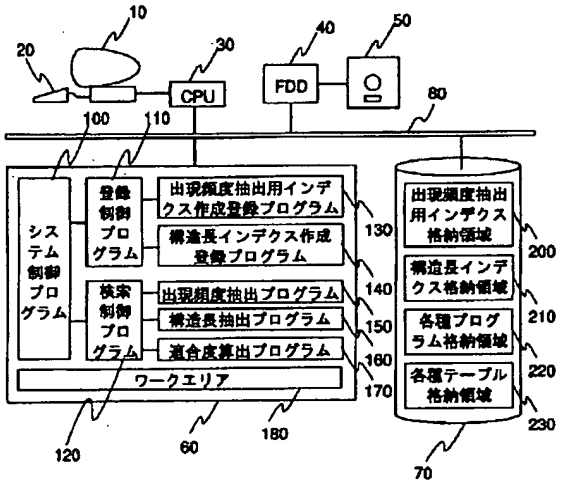
【図２２】本発明の実施例における構造長抽出プログラムの処理フローを示す図である。

【符号の説明】

１０　ディスプレイ
２０　キーボード
３０　中央演算処理装置ＣＰＵ
４０　フロッピディスクドライバ
５０　フロッピディスク
６０　主メモリ
７０　磁気ディスク
８０　バス
１００　システム制御プログラム
１１０　登録制御プログラム
１２０　検索制御プログラム
１３０　出現頻度抽出用インデクス作成登録プログラム
１４０　構造長インデクス作成登録プログラム
１５０　出現頻度抽出プログラム
１６０　構造長抽出プログラム
１７０　適合度算出プログラム
１８０　ワークエリア
２００　出現頻度抽出用インデクス格納領域
２１０　構造長インデクス格納領域
２２０　各種プログラム格納領域
２３０　各種テーブル格納領域

【図１】

図１　第一の実施例の構成図



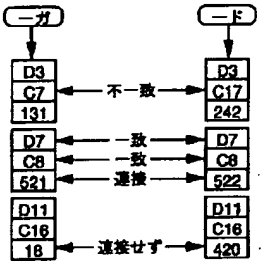
【図４】

図４　構造化全文データの例

文書識別子	文脈識別子	内容文字列
D1	C1	"SGML文書変換言語の開発とその適用事例"
	C2	"神奈川一郎"
	C3	"横浜二郎"
	C4	"川崎三郎"
	C5	"1998年10月23日"
	C6	"はじめに…"
	C7	"文書記述言語にSGMLを用いるこのによって…"
	C8	"作成したSGML文書をさまざまな…"
	C15	"適用事例"
	C16	"背景"
	C17	"現在、ISOでは…"
	C129	"変換処理の実例"
	C130	"数式の変換"
	C131	"JIS規格DTDでは、基本的には数式を…"
	C132	"ただし、行列式の場合には…"

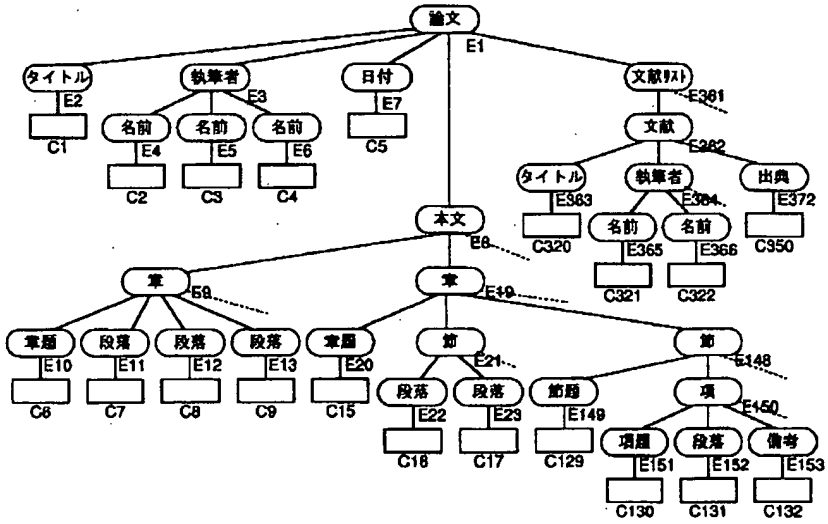
【図６】

図６　公知例２における検索処理例



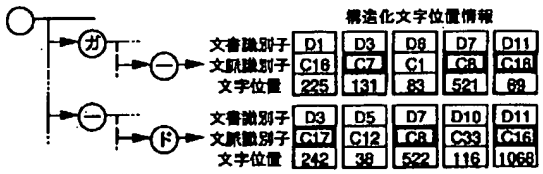
【図３】

図３　図２に示す構造化文書の論理構造



【図５】

図５　公知例２における検索用インデクスの例



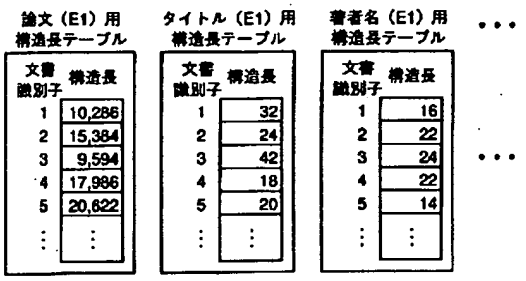
【図２】

図２　構造化文書の例

```
<!DOCTYPE 論文 SYSTEM "ronbun.dtd">
<論文>
<タイトル>SGMLにおける文書変換言語の開発とその適用事例</タイトル>
<執筆者>
<名前>神奈川一郎</名前>
<名前>横浜二郎</名前>
<名前>川崎三郎</名前>
</執筆者>
<日付>1996年10月23日</日付>
<本文>
<章>
<章題>はじめに</章題>
<段落>文書記述にSGMLを用いることによって・・・</段落>
<段落>作成したSGML文書をさまざまな・・・</段落>
</章>
<章>
<章題>適用事例</章題>
<節>
<節題>背景</節題>
<段落>現在、ISOでは・・・</段落>
</節>
<節>
<節題>変換処理の実例</節題>
<項>
<項題>数式の変換</項題>
<段落>JIS規格DTDでは、基本的に数式を・・・</段落>
<備考 type=注釈>ただし、行列式の場合には</備考>
</項>
</節>
</章>
</本文>
<文献リスト>
<文献>
<タイトル>SGMLインスタンスの変換方式の検討</タイトル>
<執筆者>
<名前>今郷詔</名前>
</執筆者>
</文献>
</文献リスト>
</論文>
```

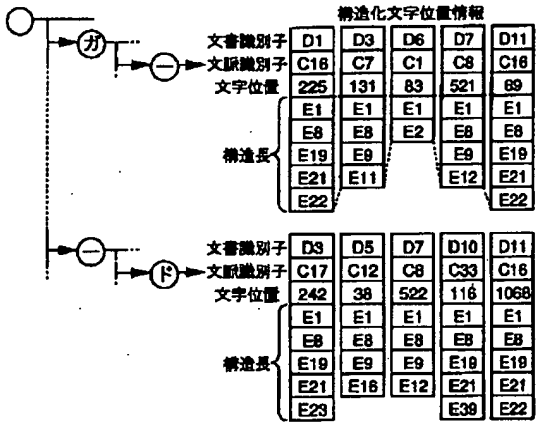
【図7】

図7　構造長テーブルを用いた構造長管理の例



【図8】

図8　検索用インデクス内に構造長を格納した時の例



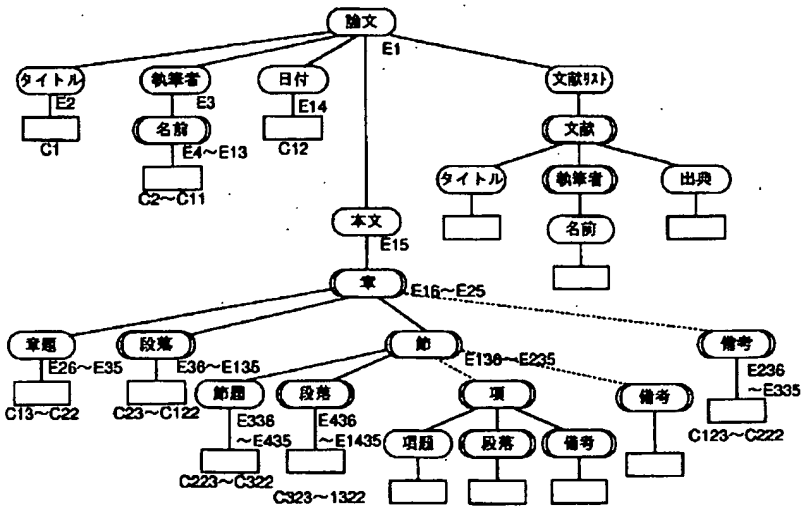
【図9】

図9 構造化文書の文書型定義（ＤＴＤ）の例

```
<!ELEMENT 論文        (ﾀｲﾄﾙ,執筆者,日付,本文,文献ﾘｽﾄ)>
<!ELEMENT ﾀｲﾄﾙ        (#PCDATA)>
<!ELEMENT 執筆者      (名前+)>
<!ELEMENT 名前        (#PCDATA)>
<!ELEMENT 日付        (#PCDATA)>
<!ELEMENT 本文        (章*)>
<!ELEMENT 章          (章題,(段落 | 節)*)>
<!ELEMENT 段落        (#PCDATA)>
<!ELEMENT 章題        (#PCDATA)>
<!ELEMENT 節          (節題,(段落)*)>
<!ELEMENT 節題        (#PCDATA)>
<!ELEMENT 文献ﾘｽﾄ     (文献+)>
<!ELEMENT 文献        (ﾀｲﾄﾙ,(執筆者+),出典)>
<!ELEMENT 出典        (#PCDATA)>
```

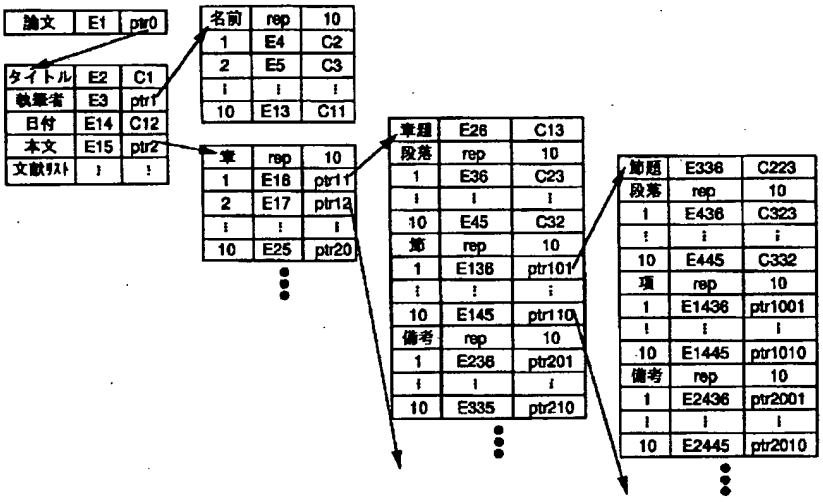
【図10】

図10　図2に示す構造化文書の論理構造



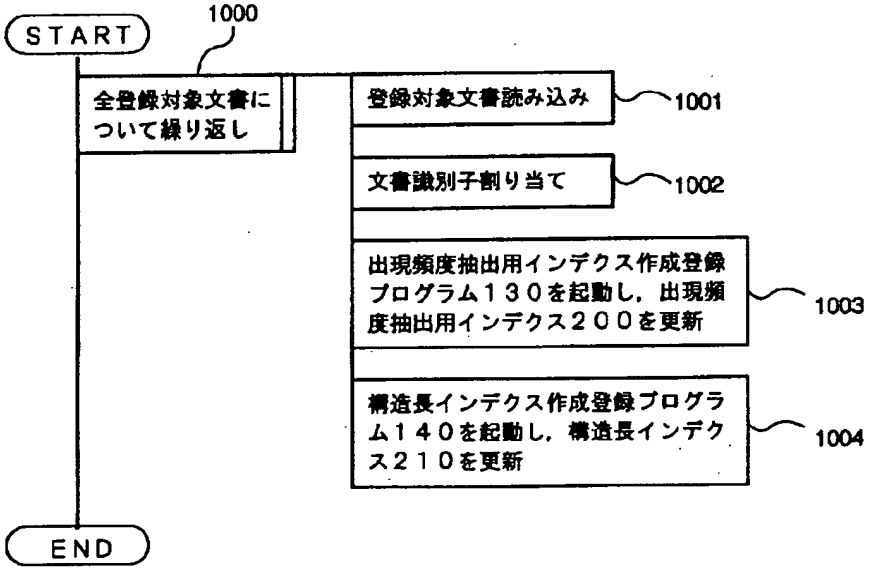
【図11】

図11　本発明第一の実施例における構造識別子管理テーブルの例



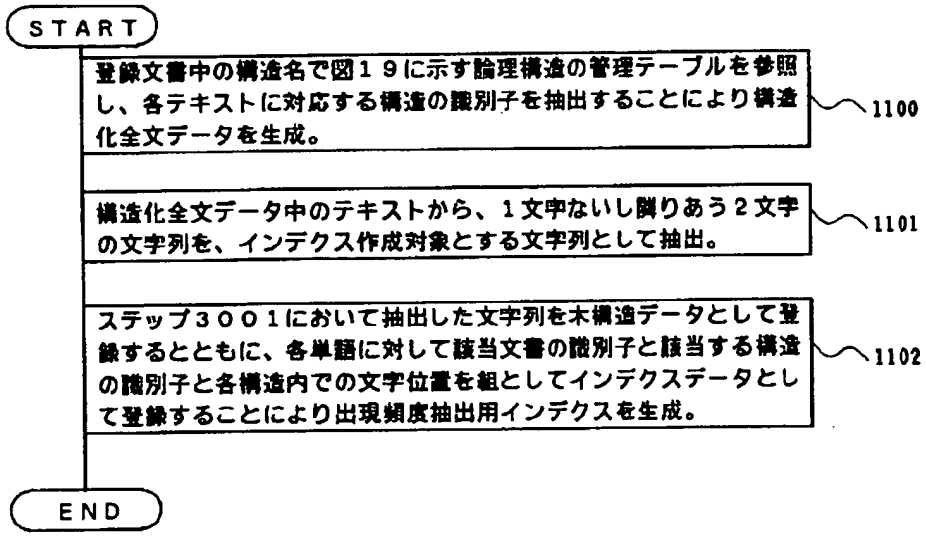
【図12】

図12　本発明第一の実施例における登録処理フロー



【図13】

図13　第一の実施例における出現頻度抽出用インデクス作成登録プログラムの処理フロー



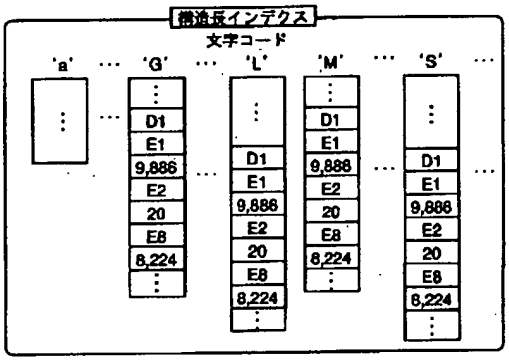
【図14】

図14　本発明第一の実施例における構造化全文データの例

文書識別子	文字列の構造識別子	内容文字列
D1	C1 (タイトル)	"SGML文書変換言語の開発とその適用事例"
	C2 (名前1)	"神奈川一郎"
	C3 (名前2)	"横浜二郎"
	C4 (名前3)	"川崎三郎"
	C12 (日付)	"1996年10月23日"
	C13 (章1-章題)	"はじめに…"
	C23 (章1-段落1)	"文書記述言語にSGMLを用いることによって…"
	C24 (章1-段落2)	"作成したSGML文書をさまざまな…"
	C14 (章2-章題)	"適用事例"
	C233 (章2-節1-節題)	"背景"
	C423 (章2-節1-段落1)	"現在、ISOでは…"
	C15 (章3-章題)	"変換処理の実例"
	C243 (章3-節1-節題)	"数式の変換"

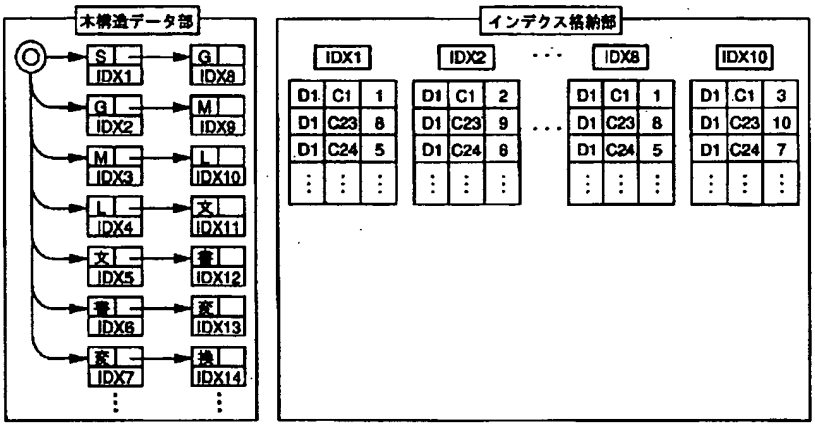
【図19】

図　19



【図15】

図15　第一の実施例における出現頻度抽出用インデクスの例



【図17】

図　17

構造別文字成分表

構造識別子	'a'	…	'G'	…	'L'	'M'	…	'S'	…	'文'	…
E1	0	…	0	…	0	0	…	0	…	0	…
E2	0	…	0	…	0	0	…	0	…	0	…
E3	0	…	0	…	0	0	…	0	…	0	…
E4	0	…	0	…	0	0	…	0	…	0	…
E5	0	…	0	…	0	0	…	0	…	0	…
E6	0	…	0	…	0	0	…	0	…	0	…
E7	0	…	0	…	0	0	…	0	…	0	…
E8	0	…	0	…	0	0	…	0	…	0	…
E9	0	…	0	…	0	0	…	0	…	0	…
E10	0	…	0	…	0	0	…	0	…	0	…
E11	0	…	0	…	0	0	…	0	…	0	…
E12	0	…	0	…	0	0	…	0	…	0	…
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

構造長リスト

構造識別子	構造長
E1	0
E2	0
E3	0
E4	0
E5	0
E6	0
E7	0
E8	0
E9	0
E10	0
E11	0
E12	0
⋮	⋮

【図18】

図　18

構造別文字成分表

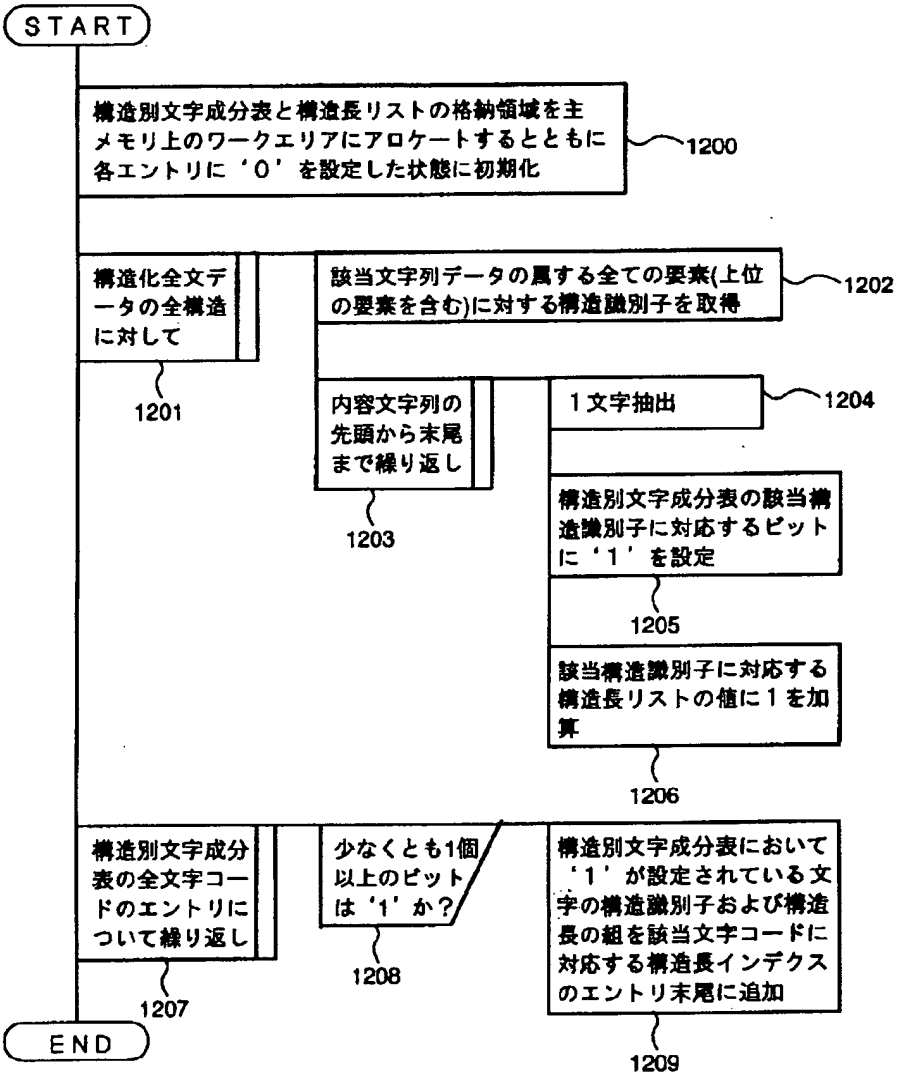
構造識別子	'a'	…	'G'	…	'L'	'M'	…	'S'	…	'文'	…
E1	0	…	1	…	1	1	…	1	…	1	…
E2	0	…	1	…	1	1	…	1	…	1	…
E3	0	…	0	…	0	0	…	0	…	0	…
E4	0	…	0	…	0	0	…	0	…	0	…
E5	0	…	0	…	0	0	…	0	…	0	…
E6	0	…	0	…	0	0	…	0	…	0	…
E7	0	…	0	…	0	0	…	0	…	0	…
E8	0	…	1	…	1	1	…	1	…	1	…
E9	0	…	1	…	1	1	…	1	…	1	…
E10	0	…	0	…	0	0	…	0	…	0	…
E11	0	…	1	…	1	1	…	1	…	1	…
E12	0	…	1	…	1	1	…	1	…	0	…
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

構造長リスト

構造識別子	構造長
E1	9,888
E2	20
E3	13
E4	5
E5	4
E6	4
E7	11
E8	8,224
E9	1,256
E10	17
E11	228
E12	188
⋮	⋮

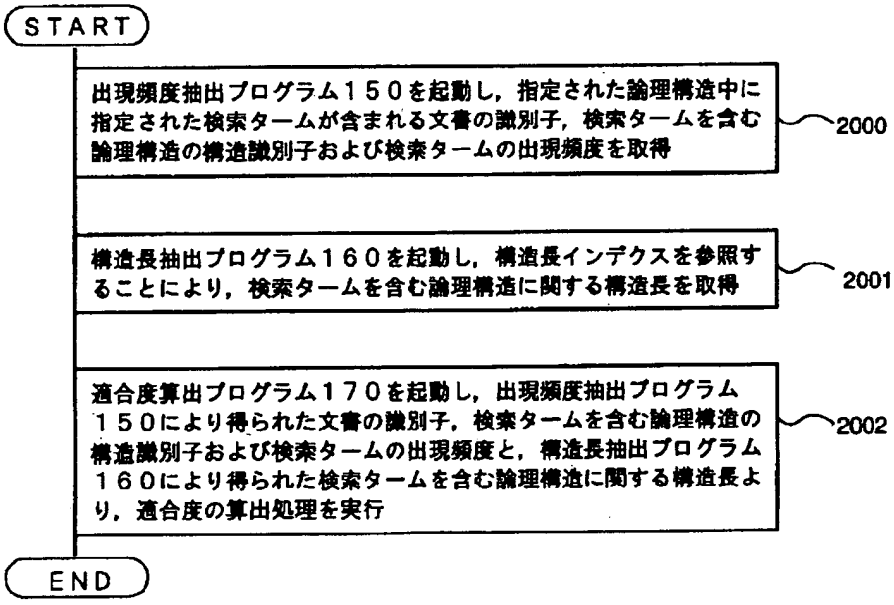
【図16】

図16　構造長インデクス作成登録プログラムの処理フロー



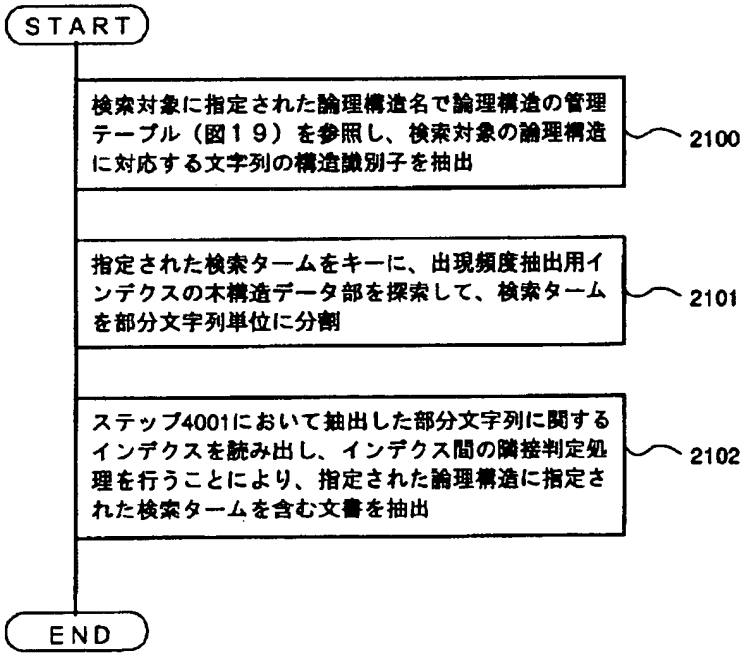
【図20】

図２０　第一の実施例における検索処理フロー



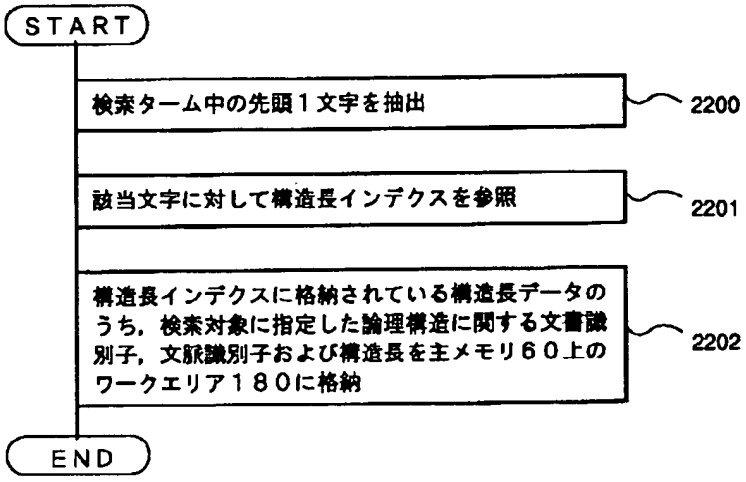
【図21】

図２１　本発明第一の実施例における出現頻度抽出プログラム１５０の処理フロー



【図22】

図22　構造長抽出プログラムの処理フロー



フロントページの続き

(72)発明者　岡本　卓哉
神奈川県川崎市幸区鹿島田890番地　株式会社日立製作所情報・通信開発本部内

(72)発明者　川下　靖司
神奈川県横浜市戸塚区戸塚町5030番地　株式会社日立製作所ソフトウェア開発本部内